

WHAT IS CLAIMED IS:

- 1 1. A method of storing page link information comprising:  
2 obtaining page link information for a set of pages, the page link information including  
3 for each page in the set a row of page identifiers of other pages;  
4 arranging the rows of page identifiers in a particular order;  
5 for each respective row:  
6 identifying a reference row, if any, that best matches the respective row in  
7 accordance with predefined row match criteria; and  
8 encoding the respective row as an identifier for the identified reference row, if  
9 any, a set of deletes representing page identifiers in the identified reference row not in the  
10 respective row, and a set of adds representing page identifiers in the respective row not in the  
11 identified reference row.
- 1 2. The method of claim 1, wherein the encoding for each respective row includes  
2 Huffman coding values representing the set of deletes and the set of adds for each respective  
3 row.
- 1 3. The method of claim 1, wherein the encoding for each respective row includes delta  
2 encoding the set of deletes and delta encoding the set of adds for each respective row.
- 1 4. The method of claim 1, wherein the encoding for each respective row includes  
2 delta encoding the set of deletes and delta encoding the set of adds for each respective  
3 row; and  
4 Huffman coding the delta encoded set of deletes and delta encoded set of adds for  
5 each respective row.
- 1 5. The method of claim 4, including  
2 sorting the page identifiers in each row in numerical order prior to performing the  
3 encoding.
- 1 6. The method of claim 5, wherein the encoding includes generating a row distance  
2 value that identifies the identified reference row and Huffman coding the row distance value.

1 7. The method of claim 4, including  
2 when no reference row exists for a respective row, encoding the respective row by  
3 encoding a null reference row identifier and a set of adds representing the page identifiers in  
4 the respective row, delta encoding the set of adds for the respective row, and Huffman coding  
5 the delta encoded set of adds for the respective row.

1 8. The method of claim 1, including  
2 when no reference row exists for a respective row, encoding the respective row by  
3 encoding a null reference row identifier and a set of adds representing the page identifiers in  
4 the respective row.

1 9. A computer program product for use in conjunction with a computer system, the  
2 computer program product comprising a computer readable storage medium and a computer  
3 program mechanism embedded therein, the computer program mechanism comprising:  
4 a first module for obtaining page link information for a set of pages, the page link  
5 information including for each page in the set a row of page identifiers of other pages; and  
6 a second module for storing the page link information, including instructions for:  
7 arranging the rows of page identifiers in a particular order;  
8 for each respective row:  
9 identifying a reference row, if any, that best matches the respective row in  
10 accordance with predefined row match criteria; and  
11 encoding the respective row as an identifier for the identified reference row, if  
12 any, a set of deletes representing page identifiers in the identified reference row not in the  
13 respective row, and a set of adds representing page identifiers in the respective row not in the  
14 identifier reference row.

1 10. The computer program product of claim 9, wherein the encoding instructions of the  
2 second module include instructions for Huffman coding values representing the set of deletes  
3 and the set of adds for each respective row.

- 1 11. The computer program product of claim 9, wherein the second module includes  
2 instructions for delta encoding the set of deletes and delta encoding the set of adds for each  
3 respective row.
- 1 12. The computer program product of claim 9, wherein the encoding instructions of the  
2 second module include instructions for delta encoding the set of deletes and delta encoding  
3 the set of adds for each respective row, and for Huffman coding the delta encoded set of  
4 deletes and delta encoded set of adds for each respective row.
- 1 13. The computer program product of claim 12, wherein the second module includes  
2 instructions for sorting the page identifiers in each row in numerical order prior to performing  
3 the encoding.
- 1 14. The computer program product of claim 13, wherein the encoding instructions of the  
2 second module include instructions for generating a row distance value that identifies the  
3 identified reference row and Huffman coding the row distance value.
- 1 15. The computer program product of claim 12, wherein the second module includes  
2 instructions, used when no reference row exists for a respective row, for encoding the  
3 respective row by encoding a null reference row identifier and a set of adds representing the  
4 page identifiers in the respective row, delta encoding the set of adds for the respective row,  
5 and Huffman coding the delta encoded set of adds for the respective row.
- 1 16. The computer program product of claim 9, wherein the second module includes  
2 instructions, used when no reference row exists for a respective row, for encoding the  
3 respective row by encoding a null reference row identifier and a set of adds representing the  
4 page identifiers in the respective row.
- 1 17. A web crawler system, comprising:  
2 a central processing unit for performing computations in accordance with stored  
3 procedures;  
4 a network interface for accessing remotely located computers via a network;

5 memory, coupled to the central processing unit, for storing procedures and data,  
6 including:  
7 a web crawler module, executable by the central processing unit, for downloading a  
8 set of pages from remotely located servers via the network interface;  
9 a first module for obtaining page link information from the set of pages, the page link  
10 information including for each page in the set a row of page identifiers of other pages; and  
11 a second module for storing the page link information, including instructions for:  
12 arranging the rows of page identifiers in a particular order;  
13 for each respective row:  
14 identifying a reference row, if any, that best matches the respective row in  
15 accordance with predefined row match criteria; and  
16 encoding the respective row as an identifier for the identified reference row, if  
17 any, a set of deletes representing page identifiers in the identified reference row not in the  
18 respective row, and a set of adds representing page identifiers in the respective row not in the  
19 identified reference row;

1 18. The system of claim 17, wherein the encoding instructions of the second module  
2 include instructions for Huffman coding values representing the set of deletes and the set of  
3 adds for each respective row.

1 19. The system of claim 17, wherein the encoding instructions of the second module  
2 include instructions for delta encoding the set of deletes and delta encoding the set of adds for  
3 each respective row.

1 20. The system of claim 17, wherein the encoding instructions of the second module  
2 includes instructions for delta encoding the set of deletes and delta encoding the set of adds  
3 for each respective row, and for Huffman coding the delta encoded set of deletes and delta  
4 encoded set of adds for each respective row.

1 21. The system of claim 20, wherein the second module includes instructions for sorting  
2 the page identifiers in each row in numerical order prior to performing the encoding.

1 22. The system of claim 21, wherein the encoding instructions of the second module  
2 include instructions for generating a row distance value that identifies the identified reference  
3 row and Huffman coding the row distance value.

1 23. The system of claim 20, wherein the second module includes instructions, used when  
2 no reference row exists for a respective row, for encoding the respective row by encoding a  
3 null reference row identifier and a set of adds representing the page identifiers in the  
4 respective row, delta encoding the set of adds for the respective row, and Huffman coding the  
5 delta encoded set of adds for the respective row.

1 24. The system of claim 17, wherein the second module includes instructions, used when  
2 no reference row exists for a respective row, for encoding the respective row by encoding a  
3 null reference row identifier and a set of adds representing the page identifiers in the  
4 respective row.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65  
66  
67  
68  
69  
70  
71  
72  
73  
74  
75  
76  
77  
78  
79  
80  
81  
82  
83  
84  
85  
86  
87  
88  
89  
90  
91  
92  
93  
94  
95  
96  
97  
98  
99  
100